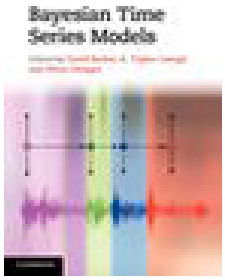


Bayesian Time Series Models

Edited by David Barber, A. Taylan Cemgil, Silvia Chiappa

Book DOI: http://dx.doi.org/10.1017/CBO9780511984679

Online ISBN: 9780511984679

Hardback ISBN: 9780521196765

Chapter

13 - Non-commutative harmonic analysis in multi-object tracking pp. 277-294

Chapter DOI: http://dx.doi.org/10.1017/CBO9780511984679.014

Cambridge University Press

13

Non-commutative harmonic analysis in multi-object tracking

Risi Kondor

13.1 Introduction

Simultaneously tracking n targets in space involves two closely coupled tasks: estimating the current positions $x_1, x_2, \ldots, x_n$ of their tracks, and estimating the assignment $\sigma\colon \{1, 2, \ldots, n\} \to \{1, 2, \ldots, n\}$ of targets to tracks. While the former is often a relatively straightforward extension of the single target case, the latter, called identity management or data association, is a fundamentally combinatorial problem, which is harder to fit in a computationally efficient probabilistic framework.

Identity management is difficult because the number of possible assignments grows with $n!$. This means that for n greater than about 10 or 12, representing the distribution $p(\sigma)$ explicitly as an array of $n!$ numbers is generally not possible.

In this chapter we discuss a solution to this problem based on the generalisation of harmonic analysis to non-commutative groups, specifically, in our case, the group of permutations. According to this theory, the Fourier transform of p takes the form

$$\widehat{p}(\lambda) = \sum_{\sigma \in \mathbb{S}_n} p(\sigma)\, \rho_\lambda(\sigma),$$

where $\mathbb{S}_n$ denotes the group of permutations of n objects, λ is a combinatorial object called an integer partition, and ρ_λ is a special matrix-valued function called a representation. These terms are defined in our short primer on representation theory in Section 13.2.

What is important to note is that, since ρ_λ is matrix-valued, each Fourier component $\widehat{p}(\lambda)$ is a matrix, not just a scalar. Apart from this surprising feature, non-commutative Fourier transforms are very similar to their familiar commutative counterparts.

In particular, we argue that there is a well-defined sense in which some of the $\widehat{p}(\lambda)$ matrices are the 'low-frequency' components of p, and approximating p with this subset of components is optimal. A large part of this chapter is focused on how to define such a notion of 'frequency', and how to find the corresponding Fourier components. We describe two seemingly very different approaches to answering this question, and find, reassuringly, that they give exactly the same answer.

Of course, in addition to a compact way of representing p, efficient inference also demands fast algorithms for updating p with observations. Section 13.6 gives an overview of the fast Fourier methods that are employed for this purpose.

13.1.1 Related work

The generalisation of harmonic analysis to non-commutative groups is based on representation theory, which, sprouting from the pioneering work of Frobenius, Schur and others at the turn of the twentieth century, has blossomed into one of the most prominent branches of algebra. The symmetric group (as the group of permutations is known) occupies a central position in this theory. For a general introduction to representation theory the reader is referred to [19], while for information on the symmetric group and its representations we recommend [18].

For much of the twentieth century, generalised Fourier transforms were the exclusive domain of pure mathematicians. It was not until the 1980s that connections to statistics and applied probability became widely recognised, thanks in particular to the work of Persi Diaconis and his collaborators. The well-known book [3] covers a wealth of topics ranging from ranking to card shuffling, and presages many of the results that we describe below, in particular with regard to spectral analysis on permutations.

Also towards the end of the 1980s a new field of computational mathematics started emerging, striving to develop fast Fourier transforms for non-commutative groups. The first such algorithm for the symmetric group is due to Clausen [2]. Later improvements and generalizations can be found in [15] and [16]. For an overview of this field, including applications, see [17].

The first context in which non-commutative harmonic analysis appeared in machine learning was multi-object tracking. This chapter is based on [12], where this idea was first introduced. Huang *et al.* [7] extended the model by deriving more general Fourier space updates, and later introduced an alternative update scheme exploiting independence [5]. The journal article [6] is a tutorial quality overview of the subject.

Besides tracking, Fourier transforms on the symmetric group can also be used to construct permutation invariant representations of graphs [11, 14], define characteristic kernels on groups [4, 9], and solve hard optimisation problems [10]. An analogue of compressed sensing for permutations is discussed in [8].

13.2 Harmonic analysis on finite groups

This section is intended as a short primer on representation theory and harmonic analysis on groups. The reader who is strictly only interested in identity management might wish to skip to Section 13.3 and refer back to this section as needed for the definitions of specific terms.

A **finite group** G is a finite set endowed with an operation $G \times G \to G$ (usually denoted multiplicatively) obeying the following axioms:

G1. For any $x, y \in G$, $xy \in G$ (closure).

G2. For any $x, y, z \in G$, $x(yz) = (xy)z$ (associativity).

G3. There is a unique $e \in G$, called the **identity** of G, such that $ex = xe = x$ for any $x \in G$.

G4. For any $x \in G$, there is a corresponding element $x^{-1} \in G$ called the **inverse** of x, such that $x x^{-1} = x^{-1} x = e$.

One important property that is missing from these axioms is commutativity, $xy = yx$. Groups that do satisfy $xy = yx$ for all x and y are called **commutative** or **Abelian** groups. A simple example of a finite commutative group is $\mathbb{Z}_n = \{0, 1, \ldots, n-1\}$, the group operation being addition modulo n. The group of permutations that appears in tracking problems, however, is not commutative.

Finite groups are quite abstract objects. One way to make them a little easier to handle is to 'model' them by square matrices that multiply the same way as the group elements do. Such a system of matrices $(\rho(x))_{x\in G}$ obeying $\rho(x)\rho(y) = \rho(xy)$ for all $x, y \in G$ is called a **representation** of G. In general, we allow representation matrices to be complex valued. Abstractly, a representation ρ is then a function $\rho\colon G \to \mathbb{C}^{d_\rho \times d_\rho}$, where d_ρ is called the **degree** or the **dimensionality** of ρ.

Once we have found one representation ρ of G, it is fairly easy to manufacture other representations. For example, if T is an invertible d_ρ-dimensional matrix, then $\rho'(x) = T^{-1}\rho(x)\,T$ is also a representation of G. Pairs of representations related to each other in this way are said to be **equivalent**.

Another way to build new representations is by taking direct sums: if ρ_1 and ρ_2 are two representations of G, then so is $\rho_1 \oplus \rho_2$, defined

$$(\rho_1 \oplus \rho_2)(x) = \rho_1(x) \oplus \rho_2(x) = \begin{pmatrix} \rho_1(x) & 0 \\ 0 & \rho_2(x) \end{pmatrix}.$$

Just as ρ' is essentially the same as ρ, $\rho_1 \oplus \rho_2$ is also not a truly novel representation. Representations which cannot be reduced into a direct sum of smaller representations (i.e., for which there is no matrix T and smaller representations ρ_1 and ρ_2 such that $\rho(x) = T^{-1}\,(\rho_1(x) \oplus \rho_2(x))\,T$ for all $x \in G$) are called **irreducible**.

A key goal in developing the representation theory of a given finite group is to find a **complete set of inequivalent irreducible representations**. We will denote such a system of representations by $\mathcal{R}_G$, and call its members **irreps** for short. Just as any natural number can be expressed as a product of primes, once we have $\mathcal{R}_G$ any representation of G can be expressed as a direct sum of irreps from $\mathcal{R}_G$, possibly conjugated by some matrix T. By a basic theorem of representation theory, if G is a finite group then $\mathcal{R}_G$ is of finite cardinality, and is well defined in the sense that if $\mathcal{R}'_G$ is a different system of irreps then there is a bijection between $\mathcal{R}_G$ and $\mathcal{R}'_G$ mapping each ρ to a ρ' with which it is equivalent. Abelian groups are special in that all their irreps are one-dimensional, so they can be regarded as just scalar functions $\rho\colon G \to \mathbb{C}$.

The concept of irreps is exactly what is needed to generalise Fourier analysis to groups. Indeed, the exponential factors $e^{-2\pi ikx}$ appearing in the discrete Fourier transform

$$\widehat{f}(k) = \sum_{x\in\{0,\ldots,n-1\}} e^{-2\pi ikx}\, f(x)$$

are nothing but the irreps of $\mathbb{Z}_n$. This suggests that the **Fourier transform** on a non-commutative finite group should be

$$\widehat{f}(\rho) = \sum_{x\in G} f(x)\,\rho(x), \qquad \rho \in \mathcal{R}_G. \tag{13.1}$$

At first sight it might seem surprising that f is a function on G, whereas $\widehat{f}$ is a sequence of matrices. It is also strange that the Fourier components, instead of corresponding to different frequencies, are now indexed by irreps. In other respects, however, Eq. (13.1) is very similar to the familiar commutative Fourier transforms. For example, we have an inverse transform

$$f(x) = \frac{1}{|G|} \sum_{\rho\in\mathcal{R}_G} d_\rho \,\mathrm{tr}\big[\rho(x)^{-1}\,\widehat{f}(\rho)\big], \tag{13.2}$$

and Eq. (13.1) also satisfies a generalised form of Parseval's theorem (more generally referred to as Plancherel's theorem), stating that with respect to the appropriate matrix norms, $f \mapsto \widehat{f}$ is a unitary transformation.

Another important property inherited from ordinary Fourier analysis is the **convolution theorem**. On a non-commutative group, the convolution of two functions f and g is defined

$$(f * g)(x) = \sum_{y \in G} f(xy^{-1})\, g(y). \tag{13.3}$$

The convolution theorem states that each component of the Fourier transform of $f * g$ is just the matrix product of the corresponding components of $\widehat{f}$ and $\widehat{g}$, that is

$$\widehat{(f * g)}(\rho) = \widehat{f}(\rho) \cdot \widehat{g}(\rho). \tag{13.4}$$

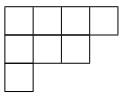
The translation and correlation theorems have similar non-commutative analogues.

13.2.1 The symmetric group

The mapping $\sigma\colon \{1, 2, \ldots, n\} \to \{1, 2, \ldots, n\}$ from targets to tracks is effectively a **permutation** of the set $\{1, 2, \ldots, n\}$. The product of two permutations is usually defined as their composition, i.e., $(\sigma_2\sigma_1)(i) = \sigma_2(\sigma_1(i))$ for all $i \in \{1, 2, \ldots, n\}$. It is easy to check that with respect to this notion of multiplication the $n!$ different permutations of $\{1, 2, \ldots, n\}$ form a non-commutative finite group. This group is called the **symmetric group** of degree n, and is denoted $\mathbb{S}_n$.

To compute the Fourier transform of the assignment distribution $p(\sigma)$, we need to study the representation theory of the symmetric group. Fortunately, starting with the pioneering work of the Rev Alfred Young at the turn of the twentieth century, mathematicians have invested a great deal of effort in exploring the representation theory of the symmetric group, and have discovered a wealth of beautiful and powerful results. Some of the questions to ask are the following: (1) How many irreps does $\mathbb{S}_n$ have and how shall we index them? (2) What is the dimensionality of each irrep ρ and how shall we index the rows and columns of $\rho(\sigma)$? (3) What are the actual $[\rho(\sigma)]_{i,j}$ matrix entries? To answer these questions Young introduced a system of combinatorial objects, which in his honour we now call Young diagrams and Young tableaux.

A **partition** of n, denoted $\lambda \vdash n$, is a k-tuple $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_k)$ satisfying $\sum_{i=1}^{k} \lambda_i = n$ and $\lambda_{i+1} \leq \lambda_i$ for $i = 1, 2, \ldots k-1$. The **Young diagram** (Ferrers diagram) of λ just consists of $\lambda_1, \lambda_2, \ldots, \lambda_k$ boxes laid down in consecutive left-justified rows. For example,



is the Young diagram of $\lambda = (4, 3, 1)$. A **Young tableau** is a Young diagram in which the boxes are bijectively filled with the numbers $1, 2, \ldots, n$, and a **standard Young tableau** is a Young tableau in which in each row the numbers increase from left to right and in each column they increase from top to bottom. For example,

1	2	5	8
3	4	7	
6			

is a standard Young tableau of shape $\lambda = (4, 3, 1)$. The set of all Young tableaux of shape λ we denote $\mathcal{T}_\lambda$.

Young discovered that there are exactly as many irreps in $\mathcal{R}_{\mathbb{S}_n}$ as there are partitions of n. Thus, instead of frequencies, in the case of the symmetric group we use partitions to index the irreps. Even more remarkably, if we employ the correct bijection between irreps and partitions, the dimensionality $d_\lambda := d_{\rho_\lambda}$ of ρ_λ is the same as the number of standard tableaux of shape λ. This suggests indexing the rows and columns of $\rho_\lambda(\sigma)$ by standard tableaux: instead of talking about the (i, j)-element of the matrix $\rho_\lambda(\sigma)$, where i and j are integers, we will talk about the (t, t')-element, where t and t' are standard tableaux of shape λ.

As regards defining the values of the actual $[\rho_\lambda(\sigma)]_{t,t'}$ matrix entries, a number of different alternatives are described in the literature, of which the most convenient one for our present purposes is **Young's orthogonal representation**, which we will abbreviate as **YOR**. In the following, whenever we refer to irreps of $\mathbb{S}_n$, we will implicitly always be referring to this system of irreducible representations.

To define YOR we need a more compact way to denote individual permutations than just writing $\sigma = [s_1, s_2, \ldots, s_n]$, where $s_1 = \sigma(1), \ldots, s_n = \sigma(n)$. The usual solution is **cycle notation**. A **cycle** $(c_1, c_2, \ldots, c_k)$ in σ is a sequence such that $\sigma(c_1) = c_2, \ldots, \sigma(c_{n-1}) = c_n$ and $\sigma(c_n) = c_1$. The cycle notation for σ consists of listing its constituent cycles, for example $\sigma = [2, 3, 1, 5, 4]$ would be written $(1, 2, 3)\,(4, 5)$. Clearly, this uniquely defines σ. Any i that are fixed by σ form single-element cycles by themselves, but these trivial cycles are usually omitted from cycle notation. The **cycle type** $\mu = (\mu_1, \mu_2, \ldots, \mu_\ell)$ of σ is the length of its cycles, listed in weakly decreasing order.

The notion of cycles and cycle type suggest some special classes of permutations. The simplest permutation is the **identity** e, which is the unique permutation of cycle type $(1, 1, \ldots, 1)$. Next, we have the class of **transpositions**, which are the permutations of cycle type $(2, 1, \ldots, 1)$. Thus, a transposition is a single two-cycle $\sigma = (i, j)$, exchanging i with j and leaving everything else fixed. **Adjacent transpositions** are special transpositions of the form $\tau_i = (i, i+1)$.

We define YOR by giving explicit formulae for the representation matrices of adjacent transpositions. Since any permutation can be written as a product of adjacent transpositions, this defines YOR on the entire group. For any standard tableau t, letting $\tau_i(t)$ be the tableau that we get from t by exchanging the numbers i and $i + 1$ in its diagram, the rule defining $\rho_\lambda(\tau_i)$ in YOR is the following: if $\tau_i(t)$ is *not* a standard tableau, then the column of $\rho_\lambda(\tau_i)$ indexed by t is zero, except for the diagonal element $[\rho_\lambda(\tau_i)]_{t,t} = 1/d_t(i, i+1)$; if $\tau_i(t)$ is a standard tableau, then in addition to this diagonal element, we also have a single non-zero off-diagonal element $[\rho_\lambda(\tau_i)]_{\tau_k(t),t} = (1 - 1/d_t(i, i+1)^2)^{1/2}$. All other matrix entries of $\rho_\lambda(\tau_i)$ are zero. In the above $d_t(i, i+1) = c_t(i+1) - c_t(i)$, where $c(j)$ is the column index minus the row index of the cell where j is located in t.

Young's orthogonal representation has a few special properties that are worth noting at this point. First, despite having stressed that representation matrices are generally complex-valued, the YOR matrices are, in fact, all real. It is a special property of $\mathbb{S}_n$ that it admits a system of irreps which is purely real. The second property is that, as the name suggests, the YOR matrices are orthogonal. In particular, $\rho_\lambda(\sigma^{-1}) = \rho_\lambda(\sigma)^\top$. Third, as is apparent from the definition, the $\rho_\lambda(\tau_i)$ matrices are extremely sparse, which will turn out to be critical for constructing fast algorithms. Finally, and this applies to all commonly used systems of irreps for $\mathbb{S}_n$, not just YOR, the representation corresponding to the partition (n) is the trivial representation $\rho_{(n)}(\sigma) = 1$ for all $\sigma \in \mathbb{S}_n$.

13.3 Band-limited approximations

Combining Eq. (13.1) with the representation theory of $\mathbb{S}_n$ tells us that the Fourier transform of the assignment distribution p is the sequence of matrices

$$\widehat{p}(\lambda) := \widehat{p}(\rho_\lambda) = \sum_{\sigma\in\mathbb{S}_n} p(\sigma)\,\rho_\lambda(\sigma), \qquad \lambda \vdash n.$$

Regarding p as a vector $\mathbf{p} \in \mathbb{R}^{\mathbb{S}_n}$, this can also be written componentwise as

$$[\widehat{p}(\lambda)]_{t,t'} = \langle \mathbf{p}, \mathbf{u}^\lambda_{t,t'} \rangle, \quad \text{where} \quad \boldsymbol{u}^\lambda_{t,t'} = \sum_{\sigma\in\mathbb{S}_n} [\rho_\lambda(\sigma)]_{t,t'}\,\mathbf{e}_\sigma,$$

and $(\mathbf{e}_\sigma)_{\sigma\in\mathbb{S}_n}$ is the canonical orthonormal basis of $\mathbb{R}^{\mathbb{S}_n}$. From this point of view, the Fourier transform is a series of projections to the subspaces

$$V_\lambda = \mathrm{span}\{\, \mathbf{u}^\lambda_{t,t'} \mid t, t' \in \mathcal{T}_\lambda \,\},$$

called the **isotypics** of $\mathbb{R}^{\mathbb{S}_n}$. By the unitarity of the Fourier transform, the isotypics are pairwise orthogonal and together span the entire space.

The key idea of this chapter is to approximate $\mathbf{p}$ by its projection to some subspace W, expressible as a sum of isotypics $W = \bigoplus_{\lambda\in\Lambda} V_\lambda$. The question is how we should choose W so as to retain as much useful information about $p(\sigma)$ as possible.

In the following we discuss two alternative approaches to answering this question. In the first approach, presented in Section 13.4, we define a Markov process governing the evolution of p, and argue that W should be the subspace least affected by stochastic noise under this model. We find that under very general conditions this subspace is indeed a sum of isotypics, specifically, in the most natural model for identity management, $W = \bigoplus_{\lambda\in\Lambda_k} V_\lambda$, where $\Lambda_k = \{\, \lambda \vdash n \mid \lambda_1 \geq n - k \,\}$. The integer parameter k plays a role akin to the cutoff frequency in low-pass filtering.

In the second approach, in Section 13.5, we ask the seemingly very different question of what sequence of subspaces $U_1, U_2, \ldots$ of $\mathbb{R}^{\mathbb{S}_n}$ capture the first-order marginals $p(\sigma(i) = j)$, second-order marginals $p(\sigma(i_1) = j_1,\ \sigma(i_2) = j_2)$, and so on, up to order k. Surprisingly, we find that the answer is again $U_k = \bigoplus_{\lambda\in\Lambda_k} V_\lambda$.

13.4 A hidden Markov model in Fourier space

Just as in tracking a single target, the natural graphical model to describe the evolution of the assignment distribution $p(\sigma)$ in identity management is a hidden Markov model. According to this model, assuming that at time t the distribution is $p_t(\sigma)$, in the absence of any observations, at time $t+1$ it will be

$$p_{t+1}(\sigma') = \sum_{\sigma\in\mathbb{S}_n} p(\sigma'|\sigma)\; p_t(\sigma), \tag{13.5}$$

where $p(\sigma'|\sigma)$ is the probability of transitioning from assignment σ to σ'. For example, if a pair of targets i_1 (assigned to track j_1) and i_2 (assigned to track j_2) come very close to each other, there is some probability that due to errors in our sensing systems their assignment might be flipped. This corresponds to transitioning from σ to $\sigma' = \tau\sigma$, where τ is the transposition (j_1, j_2).

When we do have an observation O at $t+1$, by Bayes' rule the update takes on the slightly more complicated form

$$p_{t+1}(\sigma') = \frac{p(O|\sigma') \sum_{\sigma\in\mathbb{S}_n} p(\sigma'|\sigma)\, p_t(\sigma)}{\sum_{\sigma''\in\mathbb{S}_n} p(O|\sigma'') \sum_{\sigma\in\mathbb{S}_n} p(\sigma''|\sigma)\, p_t(\sigma)}.$$

As an example, if we observe target i at track j with probability π, then

$$p(O|\sigma') = \begin{cases} \pi & \text{if } \sigma(i) = j, \\ (1-\pi)/(n-1) & \text{if } \sigma(i) \neq j. \end{cases} \tag{13.6}$$

Generally, in identity management we are interested in scenarios where observations are relatively infrequent, or the noise introduced by the transition process is relatively strong. Hence, the natural criterion for choosing the right form of band-limiting should be stability with respect to Eq. (13.5).

13.4.1 A random walk on $\mathbb{S}_n$

Equation (13.5) describes a random walk on $\mathbb{S}_n$ with transition matrix $P_{\sigma',\sigma} = p(\sigma'|\sigma)$. In particular, starting from an initial distribution $\mathbf{p}_0$, in the absence of observations, after t time steps the assignment distribution will be

$$\mathbf{p}_t = P^t \mathbf{p}_0. \tag{13.7}$$

As for random walks in general, the evolution of this process is governed by the spectral structure of P. Assuming that P is symmetric and its eigenvalues are $\alpha_1 \geq \alpha_2 \geq \ldots \geq \alpha_{n!} \geq 0$ with corresponding orthonormal eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{n!}$ and $\mathbf{p}_0 = \sum_{\mathrm{i}=1}^{\mathrm{n!}} \mathbf{p}_0^{(\mathrm{i})} \mathbf{v}_{\mathrm{i}}$, at time t,

$$\mathbf{p}_t = \sum_{i=1}^{n!} \alpha_i^t\, p_0^{(i)}\, \mathbf{v}_i. \tag{13.8}$$

Clearly, the modes of $\mathbf{p}$ corresponding to low values of α will rapidly decay. To make predictions about $\mathbf{p}$, we should concentrate on the more robust, high α modes. Hence, ideally, the approximation subspace W should be spanned by these components.

In most cases we of course do not know the exact form of P. However, there are some general considerations that can still help us find W. First of all, it is generally reasonable to assume that the probability of a transition $\sigma \mapsto \sigma'$ should only depend on σ' relative to σ. In algebraic terms, letting $\sigma' = \tau\sigma$, $p(\tau\sigma|\sigma)$ must only be a function of τ, or equivalently, $p(\sigma'|\sigma) = q(\sigma\sigma^{-1})$ for some function $q\colon \mathbb{S}_n \to \mathbb{R}$. Plugging this into Eq. (13.5) gives

$$p_{t+1}(\sigma') = \sum_{\sigma\in\mathbb{S}_n} q(\sigma'\sigma^{-1})\, p_t(\sigma),$$

which is exactly the convolution of p_t with q, as defined in Eq. (13.3). Thus, by Eq. (13.4), in Fourier space $\widehat{p}_{t+1}(\lambda) = \widehat{q}(\lambda) \cdot \widehat{p}_t(\lambda)$, and, in particular,

$$\widehat{p}_t(\lambda) = \widehat{q}(\lambda)^t \cdot \widehat{p}_0(\lambda). \tag{13.9}$$

Thus, the Fourier transform effectively block-diagonalises Eq. (13.7). From a computational point of view this is already very helpful: raising the $\widehat{q}(\lambda)$ matrices to the tth power is much cheaper than doing the same to the $n!$-dimensional P.

13.4.2 Relabelling invariance

Continuing the above line of thought, $q(\tau)$ must not depend on how we choose to label the tracks. More explicitly, if prior to a transition $\sigma \mapsto \tau\sigma$ we relabel the tracks by permuting their labels by some $\mu \in \mathbb{S}_n$, then apply τ, and finally apply μ^{-1} to restore the original labelling, then the probability of this composite transition should be the same as that of τ, i.e.,

$$q(\mu^{-1}\tau\mu) = q(\tau) \qquad \forall \mu \in \mathbb{S}_n. \tag{13.10}$$

Expressing the left-hand side by the inverse Fourier transform (13.2) and using the orthogonality of YOR gives

$$\begin{aligned} q(\mu^{-1}\tau\mu) &= \frac{1}{n!}\sum_\lambda d_\lambda \,\mathrm{tr}\,[\,\rho_\lambda(\mu^{-1}\tau^{-1}\mu)\cdot \widehat{q}(\lambda)\,] \\ &= \frac{1}{n!}\sum_\lambda d_\lambda \,\mathrm{tr}\,[\,\rho_\lambda(\mu^{-1})\cdot\rho_\lambda(\tau^{-1})\cdot\rho_\lambda(\mu)\cdot \widehat{q}(\lambda)\,] \\ &= \frac{1}{n!}\sum_\lambda d_\lambda \,\mathrm{tr}\,[\,\rho_\lambda(\tau^{-1})\cdot\rho_\lambda(\mu)\cdot \widehat{q}(\lambda)\cdot\rho_\lambda(\mu)^\top\,]. \end{aligned}$$

It is relatively easy to see that for this to equal

$$q(\tau) = \frac{1}{n!}\sum_\lambda d_\lambda \,\mathrm{tr}\,[\,\rho_\lambda(\tau^{-1})\cdot \widehat{q}(\lambda)\,]$$

for all τ and μ, we must have $T^\top \widehat{q}(\lambda)\, T = \widehat{q}(\lambda)$ for all orthogonal matrices T, which in turn implies that each $\widehat{q}(\lambda)$ is a multiple of the identity. This result is summarised in the following theorem.

Theorem 13.1 *If the transition probabilities $p(\sigma'|\sigma) = q(\sigma'\sigma^{-1})$ are relabelling-invariant in the sense of Eq. (13.10), then the Fourier transform of q is of the form*

$$\widehat{q}(\lambda) = q_\lambda I_{d_\lambda}, \qquad \lambda \vdash n,$$

where $(q_\lambda)_{\lambda\vdash n}$ are scalar coefficients and I_{d_λ} denotes the d_λ-dimensional identity matrix.

Theorem 13.1 puts a very severe restriction on the form of q. Plugging into Eq. (13.9) it tells us that in Fourier space the equation governing our random walk is simply

$$\widehat{p_t}(\lambda) = q_\lambda^t\, \widehat{p_0}(\lambda).$$

At a more abstract level, Theorem 13.1 establishes a connection between the different subspaces of $\mathbb{R}^n$ corresponding to the different Fourier components (the isotypics) and the eigenspectrum of P.

Theorem 13.2 *If $p(\sigma'|\sigma) = q(\sigma'\sigma^{-1})$ is relabelling-invariant, then the eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{n!}$ of P can be re-indexed by $\{\lambda \vdash n\}$ and $i = 1, 2, \ldots, d_\lambda^2$ so that $\{\mathbf{v}_i^\lambda\}_{i=1}^{d_\lambda^2}$ all share the same eigenvalue q_λ, and together span the isotypic V_λ.*

Hence, the different 'modes' of $\mathbf{p}$ referred to in connection with Eq. (13.8) are exactly its Fourier components! In this sense, approximating p by retaining its high q_λ Fourier components is an optimal approximation, just as in ordinary Fourier analysis low-pass filtering is optimal in the presence of high frequency noise.

13.4.3 Walks generated by transpositions

To find which Fourier components have high q_λ values, we need to be more specific about our random walk. In particular, we make the observation that while in a given finite interval of time many different subsets of targets and tracks may get exchanged, in most real-world tracking scenarios it is reasonable to assume that by making the interval between subsequent time steps sufficiently short, in each interval at most a single pair of targets will get swapped. Thus, there is no loss in restricting the set of allowable transitions to just single transpositions. Since any two transpositions τ_1 and τ_2 are related by $\tau_2 = \mu^{-1}\tau_1\mu$ for some μ, by relabelling invariance the probability of each transposition is the same, reducing the random walk to

$$p(\sigma'|\sigma) = \begin{cases} \beta & \text{if } \sigma' = (i,j)\cdot\sigma \text{ for some } 1 \le i < j \le n, \\ 1 - \binom{n}{2}\beta & \text{if } \sigma' = \sigma, \\ 0 & \text{otherwise,} \end{cases}$$

governed by the single (generally small) scalar parameter β. Now, by the argument leading to Theorem 13.1, we know that $\sum_{1\le i<j\le n}\rho_\lambda((i,j))$ is a multiple of the identity, in particular,

$$\sum_{1\le i<j\le n}\rho_\lambda((i,j)) = \frac{1}{d_\lambda}\sum_{1\le i<j\le n}\operatorname{tr}\left[\rho_\lambda((i,j))\right]\, I_{d_\lambda}.$$

In general, the function $\chi_\lambda(\sigma) = \operatorname{tr}\left[\rho_\lambda(\sigma)\right]$ is called a **character** of $\mathbb{S}_n$, and obeys

$$\chi_\lambda(\mu^{-1}\sigma\mu) = \operatorname{tr}\left[\rho_\lambda(\mu^{-1})\cdot\rho_\lambda(\sigma)\cdot\rho_\lambda(\mu)\right] = \operatorname{tr}\left[\rho_\lambda(\sigma)\cdot\rho_\lambda(\mu)\cdot\rho_\lambda(\mu)^{-1}\right] = \operatorname{tr}\left[\rho_\lambda(\sigma)\right] = \chi_\lambda(\sigma)$$

for any μ and σ. Hence, $\chi_\lambda(\tau)$ is the same for all transpositions τ, and choosing $(1,2)$ as the archetypal transposition, we can write

$$\sum_{1\le i<j\le n}\rho_\lambda((i,j)) = \binom{n}{2}\frac{\chi_\lambda((1,2))}{d_\lambda}\, I_{d_\lambda}.$$

Plugging into the Fourier transform and using the fact that for the identity permutation e, $\rho_\lambda(e) = I_{d_\lambda}$ for all λ yields that

$$q_\lambda = 1 - \beta\binom{n}{2}\left(1 - \frac{\chi_\lambda((1,2))}{d_\lambda}\right).$$

This type of expression appears in various discussions of random walks over $\mathbb{S}_n$, and, as derived in [3], it may be written explicitly as

$$q_\lambda = 1 - \beta\tbinom{n}{2}(1 - r(\lambda)) \quad \text{where} \quad r(\lambda) = \tbinom{n}{2}^{-1}\sum_i \tbinom{\lambda_i}{2} - \tbinom{\lambda'_i}{2}, \tag{13.11}$$

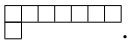
where λ' is the transpose of λ, which we get by flipping its rows and columns.

In general, we find that q_λ is highest for 'flat' partitions, which have all their squares concentrated in the first few rows. The exact order in which q_λ drops starts out as follows:

$$1 = q_{(n)} \ge q_{(n-1,1)} \ge q_{(n-2,2)} \ge q_{(n-2,1,1)} \ge q_{(n-3,3)} \ge q_{(n-3,2,1)} \ge q_{(n-3,1,1,1)} \ge q_{(n-4,4)} \ge \dots \tag{13.12}$$

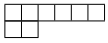
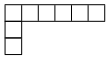
For a principled band-limited approximation to p we cut off this sequence at some point and only retain the Fourier matrices of p up to that point. It is very attractive that we can freely choose where that cutoff should be, establishing an optimal compromise between accuracy and computational expense.

Unfortunately, this freedom is somewhat limited by the fact that the dimensionality of the representations (and hence, of the Fourier matrices) grows very steeply as we move down the sequence (13.12). As we mentioned, $\rho_{(n)}$ is the trivial representation, which is one-dimensional. To find the dimensionality of the next representation, $\rho_{(n-1,1)}$, we must consider all standard tableaux of shape



Here and in the following we draw Young diagrams as if $n = 8$, but it should be understood that it is the general pattern that matters, not the exact number of boxes. In standard tableaux of the above shape, any of the numbers $2, 3, \ldots, n$ can occupy the single box in the second row. Once we have chosen this number, the rest of the standard tableau is fully determined by the 'numbers increase from left to right and top to bottom' rule. Hence, in total, there are $n-1$ standard tableaux of this shape, so $\widehat{p}((n-1,1))$ is an $n-1$-dimensional matrix.

Similarly, standard tableaux of shapes

 and 

are determined by the numbers that occupy the two boxes in the second (and third) rows, so there are $O(n^2)$ standard tableaux of each of these two shapes.

In general, the number of standard tableaux of a given shape is given by the so-called **hook rule** (see, e.g., [18]), stating that

$$d_\lambda = \frac{n!}{\prod_b \ell(b)},$$

where the product extends over all boxes of the diagram, and $\ell(b)$ is the number of boxes to the right of b plus the number of boxes beneath b plus one. The dimensionalities given by this formula for the first few partitions in the sequence (13.12) are displayed in Table 13.1.

More important than the actual d_λ values in the table is the observation that in general, d_λ grows with $n^{n-\lambda_1}$. Thus, in practice, it makes sense to cut off the Fourier expansion after

(a) the first two Fourier components $\{\widehat{p}_{(n)}, \widehat{p}_{(n-1,1)}\}$; or

(b) the first four Fourier components $\{\widehat{p}_{(n)}, \widehat{p}_{(n-1,1)}, \widehat{p}_{(n-2,2)}, \widehat{p}_{(n-2,1,1)}\}$; or

λ		d_λ
(n)		1
$(n-1,1)$		$n-1$
$(n-2,2)$		$\frac{n(n-3)}{2}$
$(n-2,1,1)$		$\frac{(n-1)(n-2)}{2}$
$(n-3,3)$		$\frac{n(n-1)(n-5)}{6}$
$(n-3,2,1)$		$\frac{n(n-2)(n-4)}{3}$

λ		d_λ
$(n-3,1,1,1)$		$\frac{(n-1)(n-2)(n-3)}{6}$
$(n-4,4)$		$\frac{n(n-1)(n-2)(n-7)}{24}$
$(n-4,3,1)$		$\frac{n(n-1)(n-3)(n-6)}{8}$
$(n-4,2,2)$		$\frac{n(n-1)(n-4)(n-5)}{12}$

Table 13.1 The size of the first few irreps of $\mathbb{S}_n$. For concreteness the diagrams are drawn as if $n = 8$.

(c) the first seven Fourier components $\{\widehat{p}_{(n)}, \widehat{p}_{(n-1,1)}, \widehat{p}_{(n-2,2)}, \widehat{p}_{(n-2,1,1)}, \widehat{p}_{(n-3,3)}, \widehat{p}_{(n-3,2,1)}, \widehat{p}_{(n-3,1,1,1)}\}$.

Going beyond these first, second and third 'order' Fourier matrices would involve $O(n^4)$-dimensional matrices, which for n in the mid teens or greater is infeasible.

13.4.4 The continuous time limit

The random walk analysis that we just presented is a somewhat simplified version of the account given in [12]. One of the differences is that the derivations in that paper were framed in terms of the **graph Laplacian**

$$\Delta_{\sigma',\sigma} = \begin{cases} -\frac{1}{\beta}\, p(\sigma'|\sigma) & \text{if } \sigma' \neq \sigma, \\ \frac{1}{\beta} \sum_{\tau \neq \sigma} p(\tau|\sigma) & \text{if } \sigma' = \sigma, \end{cases}$$

of the weighted graph corresponding to the random walk. The transition matrix P is expressed in terms of the graph Laplacian as $P = I - \beta\Delta$. In particular, for the transposition-induced random walk of Section 13.4.3,

$$\Delta_{\sigma',\sigma} = \begin{cases} -1 & \text{if } \sigma' = (i,j)\cdot\sigma \text{ for some } 1 \leq i < j \leq n, \\ \binom{n}{2} & \text{if } \sigma' = \sigma, \\ 0 & \text{otherwise.} \end{cases}$$

The eigenvalues and eigenvectors of the graph Laplacian are also referred to as the spectrum of the corresponding graph.

In general, given a subset S of a finite group G, the graph with vertex set G in which x and y are adjacent if and only if $x^{-1}y \in S$ is called the **Cayley graph** of G generated by S. Thus, by our earlier results, we have the following theorem.

Theorem 13.3 *The eigenvalues of the Cayley graph of $\mathbb{S}_n$ generated by transpositions are*

$$\alpha_\lambda = \binom{n}{2}\left(1 - \frac{\chi_\lambda((1,2))}{d_\lambda}\right) = \binom{n}{2}(1 - r(\lambda)), \qquad \lambda \vdash n,$$

where $r(\lambda)$ is defined as in Eq. (13.11), and each α_λ is d_λ^2-fold degenerate.

In recent years spectral graph theory has become popular in machine learning in a variety of contexts from dimensionality reduction [1], through constructing kernels [13], to semi-supervised learning. The Laplacian of the Cayley graph establishes a connection to this literature. A detailed account of kernels on finite groups is given in [9], and [4] investigates the properties of kernels on groups in general.

An important advantage of the Laplacian formulation is that it lets us take the continuous time limit of the random walk. Dividing the interval from t to $t+1$ into k equal time steps,

$$\mathbf{p}_{t+1} = \left(I - \frac{\beta\Delta}{k}\right)^k \mathbf{p}_t.$$

In the limit $k \to \infty$, where in any finite interval of time there are an infinite number of opportunities for a given transition to take place, but the probability of it taking place in any specific infinitesimal sub-interval is infinitesimally small, the expression in parentheses

becomes the **matrix exponential** $e^{-\beta\Delta}$, and we arrive at the equation describing **diffusion** on our graph,

$$\mathbf{p}'_t = e^{-(t'-t)\beta\Delta}\,\mathbf{p}(t), \tag{13.13}$$

where t and t' are now real numbers. By analogy with Eq. (13.9),

$$\widehat{p_{t'}}(\lambda) = e^{-\alpha_\lambda\beta(t'-t)}\,\widehat{p_t}(\lambda),$$

so in Fourier space diffusion just amounts to rescaling the $\widehat{p}(\lambda)$ Fourier matrices.

In most real-world scenarios transitions happen in continuous time, so the diffusion model is, in fact, more appropriate than the discrete time random walk, and this is the model that we implemented in our experiments.

13.5 Approximations in terms of marginals

The random walk analysis of the previous section is mathematically compelling, but sheds no light on what information is captured by the different Fourier components. Leaving the Fourier formalism aside for the moment, let us ask what other, more intuitive ways we could find an appropriate subspace W for approximating $\mathbf{p}$.

One traditional approach to identity management is to just keep track of the $n\times n$ matrix of probabilities

$$M^{(1)}_{j,i} = \text{Prob}\,(\text{ target } i \text{ is assigned to track } j\,) = \sum_{\sigma(i)=j} p(\sigma).$$

Clearly, this is a very impoverished representation for p, but it does have the merit of being fast to update. Huang *et al.* [6] demonstrate the limitations of such a first-order approach on a specific example. A more refined approach is to look at the $n(n-1)$-dimensional matrix of second-order marginals

$$M^{(2)}_{(j_1,j_2),(i_1,i_2)} = \text{Prob}\,(\,i_1 \text{ is at } j_1 \text{ and } i_2 \text{ is at } j_2\,) = \sum_{\sigma(i_1)=j_1,\,\sigma(i_2)=j_2} p(\sigma),$$

and so on, to higher orders. In general, kth order marginals can be expressed as an inner product

$$M^{(k)}_{(j_1,\ldots,j_k),(i_1,\ldots,i_k)} = \langle \mathbf{p}, \mathbf{u}_{(i_1,\ldots,i_k),(j_1,\ldots,j_k)}\rangle,$$

where $\mathbf{u}_{(i_1,\ldots,i_k),(j_1,\ldots,j_k)} = \sum_{\sigma(i_1)=j_1,\ldots,\sigma(i_k)=j_k} \mathbf{e}_\sigma$, and representing $p(\sigma)$ by $M^{(k)}$ corresponds to approximating it by its projection to

$$U_k = \text{span}\left\{\, \mathbf{u}_{(i_1,\ldots,i_k),(j_1,\ldots,j_k)} \mid i_1,\ldots,i_k,j_1,\ldots,j_k \in \{1,2,\ldots,n\}\,\right\}. \tag{13.14}$$

The natural question to ask is how the hierarchy of subspaces $U_1 \subset U_2 \subset \ldots \mathbb{R}^{\mathbb{S}_n}$ is related to the $(V_\lambda)_{\lambda\vdash n}$ isotypics. To answer this, the first thing to note is that in matrix form

$$M^{(1)} = \sum_{\sigma\in\mathbb{S}_n} p(\sigma)\,P^{(1)}(\sigma),$$

where $P^{(1)}(\sigma)$ are the usual permutation matrices

$$[P^{(1)}(\sigma)]_{j,i} = \begin{cases} 1 & \text{if } \sigma(i) = j, \\ 0 & \text{otherwise.} \end{cases}$$

Similarly, the matrix of kth order marginals can be written as

$$M^{(k)} = \sum_{\sigma \in \mathbb{S}_n} p(\sigma)\, P^{(k)}(\sigma), \tag{13.15}$$

where $P^{(k)}$ is the kth order permutation matrix

$$[P^{(k)}(\sigma)]_{(j_1,\ldots,j_k),(i_1,\ldots,i_k)} = \begin{cases} 1 & \text{if } \sigma(i_1) = j_1,\ \sigma(i_2) = j_2,\ \ldots,\ \sigma(i_k) = j_k, \\ 0 & \text{otherwise,} \end{cases}$$

which is $n!/(n-k)!$-dimensional.

Such generalised permutation matrices map the basis vector labelled $(i_1, i_2, \ldots, i_k)$ to the basis vector labelled $(\sigma(i_1), \ldots, \sigma(i_k))$. It follows that $P^{(k)}(\sigma_2)P^{(k)}(\sigma_1)$ maps $(i_1, i_2, \ldots, i_k)$ to $(\sigma_2\sigma_1(i_1), \ldots, \sigma_2\sigma_1(i_k))$. In other words, $P^{(k)}$ is a representation of $\mathbb{S}_n$, and hence it must be expressible as a sum of irreps in the form

$$P^{(k)}(\sigma) = T_k^{-1} \Big[\bigoplus_{\lambda \vdash n} \bigoplus_{m=1}^{K_{k,\lambda}} \rho_\lambda(\sigma) \Big] T_k \qquad \forall\, \sigma \in \mathbb{S}_n,$$

for some appropriate choice of multiplicities $K_{k,\lambda}$ and invertible matrix T_k (if a particular irrep does not feature in this sum, then we just set the corresponding $K_{k,\lambda}$ to zero). Plugging this into Eq. (13.15) expresses $M^{(k)}$ directly in terms of the Fourier components of p as

$$M^{(k)} = T_k^{-1} \Big[\bigoplus_{\lambda \vdash n} \bigoplus_{m=1}^{K_{k,\lambda}} \widehat{p}(\lambda) \Big] T_k,$$

implying that the subspace of kth order marginals is the sum of isotypics

$$U_k = \bigoplus_{\substack{\lambda \vdash n \\ K_{k,\lambda} \geq 1}} V_\lambda.$$

The general answer to what the $K_{k,\lambda}$ and T are is given by a result called James' Submodule Theorem, as explained in [6]. Stating the theorem verbatim would require introducing additional notation and terminology. Instead, we just state its specialisation to the case of interest to us.

Theorem 13.4 *The space (13.14) of kth order marginals is the direct sum of isotypics*

$$U_k = \bigoplus_{\substack{\lambda \vdash n \\ \lambda_1 \geq n-k}} V_\lambda.$$

Thus, the intuitive notion of approximating p by its first, second, third, etc. order marginals leads to exactly the same approximation as the random walk analysis did. From this point of view, which is discussed extensively in [3, 6], and elsewhere, the significance of the Fourier formalism is that it provides a canonical basis for the U_k subspaces, eliminating the otherwise non-trivial linear dependencies between marginals. In addition, as we shall see in the next section, the structure of the Fourier transform is also the key to devising fast algorithms for updating p with observations.

13.6 Efficient computation

The previous two sections have made a strong case for approximating p in a particular way, by discarding all but a small number of specific Fourier components. A compact way to store p is only one half of the story, however: if any of the computations required to run the hidden Markov model demanded full Fourier transforms, our approach would still be infeasible. At a minimum, we need to be able to efficiently perform the following three operations:

1. **Rollup**, which is updating p between observations by the noise model, as expressed in Eq. (13.13).
2. **Conditioning**, which is the word used for updating p with observations, such as Eq. (13.6).
3. **Prediction**, which typically involves returning the maximum a posteriori permutation, or computing some set of marginals, such as $p_i(j) = p(\sigma(i) = j)$.

Each of these operations is to be applied to the kth order band-limited approximation described in the previous sections, consisting of the Fourier components

$$\widehat{p}(\lambda), \qquad \lambda \in \Lambda_k = \{\, \lambda \vdash n \mid \lambda_1 \geq n - k \,\}.$$

As we have seen, the largest of these matrices are $O(n^k)$-dimensional, so the total storage complexity is $O(n^{2k})$. We assume that at time zero the correct assignment is known, and that without loss of generality it is the identity permutation, so we initialise our model with $\widehat{p}(\lambda) = I_{d_\lambda}$, since $\rho_\lambda(e) = I_{d_\lambda}$. An alternative way to seed the model would be to set $\widehat{p}((n)) = 1$ and $\widehat{p}(\lambda) = 0$ for all $\lambda \neq (n)$, corresponding to the uniform distribution.

Of the three operations above, rollup is very easy to perform in Fourier space, since as we have seen, it just corresponds to rescaling the individual Fourier matrices according to $\widehat{p}_{t'}(\lambda) = e^{-\alpha_\lambda \beta (t'-t)}\, \widehat{p}_t(\lambda)$. This takes only $O(n^{2k})$ time.

Deriving algorithms for conditioning and prediction that run similarly fast is somewhat more involved, and requires considering projections of $\mathbf{p}$ to yet another system of subspaces, namely

$$R_{(i_1,\ldots,i_\ell),(j_1,\ldots,j_\ell)} = \operatorname{span}\{\, \mathbf{e}_\sigma \mid \sigma(i_1) = j_1, \ldots, \sigma(i_\ell) = j_\ell \,\}, \quad i_1 < i_2 < \ldots < i_e$$

if we are interested in conditioning on or predicting marginals up to order ℓ. Clearly, for any choice of ℓ and $i_1, \ldots, i_\ell$,

$$\mathbb{R}^{\mathbb{S}_n} = \bigoplus_{j_1,\ldots,j_\ell} R_{(i_1,\ldots,i_\ell),(j_1,\ldots,j_\ell)},$$

where the sum extends over all choices of mutually distinct $j_1, \ldots, j_\ell$. Moreover, $\{\, \sigma \mid \sigma(i_1) = j_1, \ldots \sigma(i_\ell) = j_\ell \,\}$ is a structure very similar to $\mathbb{S}_{n-\ell}$ (technically, it is an $\mathbb{S}_{n-\ell}$-coset), since it is a set of $(n-\ell)!$ permutations that map any i which is not one of $i_1, \ldots, i_\ell$ to any position j, as long as it is not $j_1, \ldots, j_\ell$. This implies that each $R_{(i_1,\ldots,i_\ell),(j_1,\ldots,j_\ell)}$ subspace has its own Fourier transform with respect to $\mathbb{S}_{n-\ell}$. Our key computational trick is to relate the individual components of these $\mathbb{S}_{n-\ell}$-transforms to the global $\mathbb{S}_n$-transform. For simplicity we only derive these relationships for the 'first-order' subspaces $R_{i,j} \equiv R_{(i),(j)}$. The higher-order relations ($\ell > 1$) can be derived by recursively applying the first-order ones.

Identifying $\mathbb{S}_{n-1}$ with the subgroup of permutations that fix n and defining $[\![a,b]\!]$ as the permutation

$$[\![a,b]\!](i) = \begin{cases} i+1 & \text{if } i = a,\dots,b-1, \\ a & \text{if } i = b, \\ i & \text{otherwise,} \end{cases}$$

any σ satisfying $\sigma(i) = j$ can be uniquely written as $\sigma = [\![j,n]\!]\,\tau\,[\![i,n]\!]^{-1}$ for some $\tau \in \mathbb{S}_{n-1}$. Thus, the projection of a general vector $p \in \mathbb{R}^{\mathbb{S}_n}$ to $R_{i,j}$ is identified with $p_{i,j} \in \mathbb{R}^{\mathbb{S}_{n-1}}$ defined $p_{i,j}(\tau) = p([\![j,n]\!]\,\tau\,[\![i,n]\!]^{-1})$. Writing the full Fourier transform as

$$\begin{aligned} \widehat{p}(\lambda) &= \sum_{j=1}^{n-1} \sum_{\tau\in\mathbb{S}_{n-1}} p([\![j,n]\!]\,\tau\,[\![i,n]\!]^{-1})\,\rho_\lambda([\![j,n]\!]\,\tau\,[\![i,n]\!]^{-1}) \\ &= \sum_{j=1}^{n-1} \rho_\lambda([\![j,n]\!]) \Big[\sum_{\tau\in\mathbb{S}_{n-1}} p([\![j,n]\!]\,\tau\,[\![i,n]\!]^{-1})\,\rho_\lambda(\tau) \Big] \rho_\lambda([\![i,n]\!])^{-1}, \end{aligned}$$

the expression in brackets is seen to be almost the same as the Fourier transform of $p_{i,j}$, except that ρ_λ is an irrep of $\mathbb{S}_n$ and not of $\mathbb{S}_{n-1}$. By complete reducibility we know that if τ is restricted to $\mathbb{S}_{n-1}$, then $\rho_\lambda(\tau)$ must be expressible as a sum of $\mathbb{S}_{n-1}$-irreps, but in general the exact form of this decomposition can be complicated. In YOR, however, it is easy to check that the decomposition is just

$$\rho_\lambda(\tau) = \bigoplus_{\lambda^-\in\lambda\downarrow_{n-1}} \rho_{\lambda^-}(\tau), \qquad \tau \in \mathbb{S}_{n-1},$$

where $\lambda\downarrow_{n-1}$ denotes the set of all partitions of $n-1$ that can be derived from λ by removing one box. Thus, $\widehat{p}$ can be computed from $(\widehat{p}_{i,j})_{j=1}^n$ by

$$\widehat{p}(\lambda) = \sum_{j=1}^{n} \rho_\lambda([\![j,n]\!]) \Big[\bigoplus_{\lambda^-\in\lambda\downarrow_{n-1}} \widehat{p}_{i,j}(\lambda^-) \Big] \rho_\lambda([\![i,n]\!])^\top. \tag{13.16}$$

A short computation shows that the inverse of this transformation is

$$\widehat{p}_{i,j}(\lambda^-) = \frac{1}{n\,d_{\lambda^-}} \sum_{\lambda\in\lambda^-\uparrow^n} d_\lambda \left[\rho_\lambda([\![j,n]\!])^\top\,\widehat{p}(\lambda)\,\rho_\lambda([\![i,n]\!]) \right]_{\lambda^-}, \tag{13.17}$$

where $\lambda^-\uparrow^n$ denotes the set of those partitions of n that we can get by adding a single box to λ, and $[M]_{\lambda^-}$ denotes the block of M corresponding to λ^-. In [12] we explain that thanks to the special structure of YOR, these computations can be performed very fast: for kth-order band-limited functions the complexity of Eqs. (13.16) and (13.17) is just $O(n^{2k+2})$. If we are only interested in a single projection $\widehat{p}_{i,j}$, then this is further reduced to $O(n^{2k+1})$. We remark that these operations are a modified form of the elementary steps in Clausen's FFT [2].

Conditioning on the assignment of individual targets and computing marginals can both be expressed in terms of the forward (13.16) and backward (13.17) transforms. For example, if at a given moment in time target i is observed to be at track j with probability π, then by Bayes' rule, p is to be updated to

$$p'(\sigma) = p(\sigma|O) = \frac{p(O|\sigma)\,p(\sigma)}{\sum_{\sigma'\in\mathbb{S}_n} p(O|\sigma')\,p(\sigma')},$$

where

$$p(O|\sigma) = \begin{cases} \pi & \text{if } \sigma(i) = j, \\ (1-\pi)/(n-1) & \text{if } \sigma(i) \neq j. \end{cases} \tag{13.18}$$

In terms of vectors this is simply $\mathbf{p}' \propto \frac{1-\pi}{n-1}\,\mathbf{p} + \frac{\pi n-1}{n-1}\,\mathbf{p}_{i\to j}$, where $\mathbf{p}_{i\to j}$ is the projection of $\mathbf{p}$ to $R_{i,j}$. Thus, the update can be performed by computing $\mathbf{p}_{i\to j}$ by Eq. (13.17), rescaling by the respective factors $\frac{1-\pi}{n-1}$ and $\frac{\pi n-1}{n-1}$, transforming back by Eq. (13.16), and finally normalising. All this can be done in time $O(n^{2k+1})$. Higher-order observations of the form $\sigma(i_1) = j_1, \ldots, \sigma(i_\ell) = j_\ell$ would involve projecting to the corresponding $R_{(i_1,\ldots,i_\ell),(j_1,\ldots,j_\ell)}$ subspace and would have the same time complexity.

Prediction in the simplest case consists of returning estimates of the probabilities $p(\sigma(i) = j)$. Computing these probabilities is again achieved by transforming to the $R_{i,j}$ subspaces. In particular, since $\rho_{(n-1)}$ is the trivial representation of $\mathbb{S}_{n-1}$, the one-dimensional Fourier component $\widehat{p}_{i,j}((n-1)) = \sum_{\tau\in\mathbb{S}_{n-1}} p_{i,j}(\tau)$ is exactly $p(\sigma(i) = j)$. In computing this single component, the sum in Eq. (13.17) need only extend over $\lambda = (n)$ and $(n-1)$, thus $p(\sigma(i) = j)$ can be computed from $\widehat{p}$ in $O(n^3)$ time. Naturally, computing $p(\sigma(i) = j)$ for all j then takes $O(n^4)$ time.

13.6.1 Truncation and positivity

In the above discussion we implicitly made the assumption that if p is initialised to be kth order band-limited, then as it evolves in time, it will preserve this structure. This is indeed true of the rollup update, but in the conditioning step adding the rescaled $\mathbf{p}_{i\to j}$ back onto $\mathbf{p}$ will generally activate additional Fourier components. Thus, conditioning must involve truncation in the Fourier domain.

To ensure that $\mathbf{p}$ still remains a probability distribution, we need to normalise and enforce pointwise positivity. Normalisation is relatively easy, since, as for $p_{i\to j}$, the total weight $\sum_{\sigma\in\mathbb{S}_n} p(\sigma)$ can be read off from $\widehat{p}((n))$. If this value strays from unity, all we need to do is divide all the $\widehat{p}(\lambda)$ matrices by it to renormalise.

Positivity is more difficult to enforce. In [12] we argued that in most cases even when $p(\sigma)$ becomes negative for some permutations, this does not seem to be a serious problem for predicting the marginals that we are ultimately interested in. An alternative approach introduced in [7], which seems to do somewhat better, is to use a quadratic program to project $\mathbf{p}$ back onto an appropriate marginal polytope after each conditioning step.

13.6.2 Kronecker conditioning

Our fast, $O(n^{2k+1})$ complexity method for conditioning in Fourier space relies heavily on the specific form (13.18) of the likelihood in our observation model. It is interesting to ask how the posterior might be computed in Fourier space if $g(\sigma) = p(O|\sigma)$ was a general function on permutations. In [6] it is shown this is related to the so called Clebsch–Gordan decomposition, which tells us how the tensor (or Kronecker) product of two representations decomposes into a direct sum:

$$C^{\dagger}_{\lambda_1,\lambda_2}\,[\rho_{\lambda_1}(\sigma)\otimes\rho_{\lambda_2}(\sigma)]\,C_{\lambda_1,\lambda_2} = \bigoplus_{\lambda\vdash n}\bigoplus_{i=1}^{z_{\lambda_1,\lambda_2,\lambda}}\rho_\lambda(\sigma), \qquad \forall\,\sigma\in\mathbb{S}_n, \tag{13.19}$$

where the $d_{\lambda_1}d_{\lambda_2}$-dimensional constant matrix C_{λ_1,λ_2}, and the $z_{\lambda_1,\lambda_2,\lambda}$ multiplicities are universal (albeit not easily computable) constants. In particular, they show that the Fourier

transform of the unnormalised posterior $p'(\sigma) = g(\sigma)\, p(\sigma)$ is

$$\widehat{p}\,'(\lambda) = \frac{1}{n!\, d_\lambda} \sum_{\lambda_1,\lambda_2 \vdash n} d_{\lambda_1} d_{\lambda_2} \sum_{i=1}^{z_{\lambda_1,\lambda_2,\lambda}} \left[C^{\dagger}_{\lambda_1,\lambda_2} \left[\widehat{g}(\lambda_1) \otimes \widehat{p}(\lambda_2) \right] C_{\lambda_1,\lambda_2} \right]_{\lambda,i}, \tag{13.20}$$

where $[\,\cdot\,]_{\lambda,i}$ corresponds to the 'ith λ-block' of the matrix in brackets according to the decomposition (13.19).

The price to pay for the generality of this formula is its computational expense: in contrast to the $O(n^{2k+1})$ complexity of conditioning with Eq. (13.18), if we assume that g is mth order band-limited, the complexity of computing Eq. (13.20) is $O(n^{3k+2m})$. Huang *et al.* [6] argue that in practice the C_{λ_1,λ_2} matrices are sparse, which somewhat reduces this computational burden, and manage to get empirical results using their approach for $n = 11$.

13.6.3 Empirical performance

Both our group and the Huang–Guestrin–Guibas group have performed experiments to validate the Fourier approach to identity management, but side-by-side comparisons with traditional algorithms are difficult for lack of standard benchmark datasets. Our group culled data from an online source of flight paths of commercial aircraft, while Huang *et al.* collected data of people moving around in a room, and later of ants in an enclosed space.

All experiments bore out the general rule that the more Fourier components that an algorithm can maintain, the better its predictions will be. Using the fast updates described above our algorithms can afford to maintain second-order Fourier components up to about $n = 30$, and third-order components up to about $n = 15$. Typically, each update takes only a few seconds on an ordinary desktop computer.

In contrast, more traditional identity management algorithms generally either store the entire distribution explicitly, which is only feasible for $n \leq 12$, or in some form work with first-order marginals. Thus, Fourier algorithms have a definite edge in the intermediate $12 < n \leq 30$ range.

Of course, all these statements relate to the scenario described in the introduction, where observations are relatively rare and p becomes appreciably spread out over many permutations. If the uncertainty can be localised to a relatively small set of permutations or a subset of the targets, then a particle filter method or the factorisation approach in [5] might be more appropriate. For more information on the experiments the reader is referred to [12] and [7].

13.7 Conclusions

Identity management is a hard problem because it involves inference over a combinatorial structure, namely the group of permutations. We argued that the right way to approach this problem is by Fourier analysis on the symmetric group.

While at first sight the Fourier transform on the symmetric group seems like a rather abstract mathematical construction, we have shown that the individual Fourier components of the assignment distribution $p(\sigma)$ can be interpreted in terms of both the modes of a random walk on permutations and in terms of the natural hierarchy of marginal probabilities. In particular, there is a sense in which certain Fourier components capture the 'low-frequency' information in $p(\sigma)$, and estimating $p(\sigma)$ in terms of these components is optimal.

In addition to discussing this principled way of approximating distributions over permutations, we also derived algorithms for efficiently updating it in a Bayesian way with observations. In general, we find that the kth order Fourier approximation to p has space complexity $O(n^{2k})$ and time complexity $O(n^{2k+1})$.

While the present chapter discussed identity management in isolation, in many real-world settings one would want to couple such a system to some other model describing the position of the individual targets, so that the position information can help disambiguate the identity information and vice versa. This introduces a variety of interesting issues, which are still the subject of research.

Acknowledgments The author thanks Andrew Howard and Tony Jebara for their contributions to [12], the original paper that this chapter is largely based on. He is also indebted to Jonathan Huang, Carlos Guestrin and Leonidas Guibas for various discussions.

Bibliography

[1] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. In *Neural Information Processing Systems (NIPS)*, 2001.

[2] M. Clausen. Fast generalized Fourier transforms. *Theoretical Computer Science*, **67**:55–63, 1989.

[3] P. Diaconis. *Group Representations in Probability and Statistics*. IMS Lecture Series. Institute of Mathematical Statistics, 1988.

[4] K. Fukumizu, B. K. Sriperumbudur, A. Gretton and B. Schölkopf. Characteristic kernels on groups and semigroups. In *Neural Information Processing Systems (NIPS)*, 2008.

[5] J. Huang, C. Guestrin, X. Jiang and L. Guibas. Exploiting probabilistic independence for permutations. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2009.

[6] J. Huang, C. Guestrin and L. Guibas. Fourier theoretic probabilistic inference over permutations. *Journal of Machine Learning Research*, **10**:997–1070, 2009.

[7] J. Huang, C. Guestrin and L. Guibas. Efficient inference for distributions on permutations. In *Neural Information Processing Systems (NIPS)*, 2007.

[8] S. Jagabathula and D. Shah. Inferring rankings under constrained sensing. In *Neural Information Processing Systems (NIPS)*, 2008.

[9] R. Kondor. *Group theoretical methods in machine learning*. PhD thesis, Columbia University, 2008.

[10] R. Kondor. A Fourier space algorithm for solving quadratic assignment problems. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2010.

[11] R. Kondor and K. M. Borgwardt. The skew spectrum of graphs. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2008.

[12] R. Kondor, A. Howard, and T. Jebara. Multi-object tracking with representations of the symmetric group. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2007.

[13] R. Kondor and J. Lafferty. Diffusion kernels on graphs and other discrete structures. In *Proceedings of the International Conference on Machine Learning (ICML) 2002*, 2002.

[14] R. Kondor, N. Shervashidze and K. M. Borgwardt. The graphlet spectrum. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2009.

[15] D. K. Maslen. The efficient computation of Fourier transforms on the symmetric group. *Mathematics of Computation*, **67**(223):1121–1147, 1998.

[16] D. K. Maslen and D. N. Rockmore. Double coset decompositions and computational harmonic analysis on groups. *Journal of Fourier Analysis and Applications*, **6**(4), 2000.

[17] D. N. Rockmore. Some applications of generalized FFTs. *Proceedings of the DIMACS workshop on groups and computation 1995*, 1997.

[18] B. E. Sagan. *The Symmetric Group. Representations, combinatorial algorithms and symmetric functions*, volume 203 of Graduate Texts in Mathematics. Springer, 2001.

[19] J.-P. Serre. *Linear Representations of Finite Groups*, volume 42 of Graduate Texts in Mathematics. Springer-Verlag, 1977.

Contributor

Risi Kondor, Center for the Mathematics of Information, California Institute of Technology, USA